Introduction to mathematical Statistics

Chapter 6. Interval Estimation

1.Introduction to the Pivotal Quantity Approach, with an application in:

Deriving the Exact Confidence Interval for the population mean when the population is normal and the population variance is known.

Example. Let $X_1, X_2, \cdots, X_n$ be a random sample from a normal population $N(\mu, \sigma^2)$

1. Find the MLE of μ
2. Please construct a 95% confidence interval for μ .

Solution.

1. $\hat{\mu} = \overline{X}$
2. $P(a \le \mu \le b) = 0.95$

How to construct a CI for μ

1. Start with the point estimator for μ

$$\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$$

2. Based on $\overline{X}$, we will construct a new R.V. that is called a **pivotal quantity** for μ .

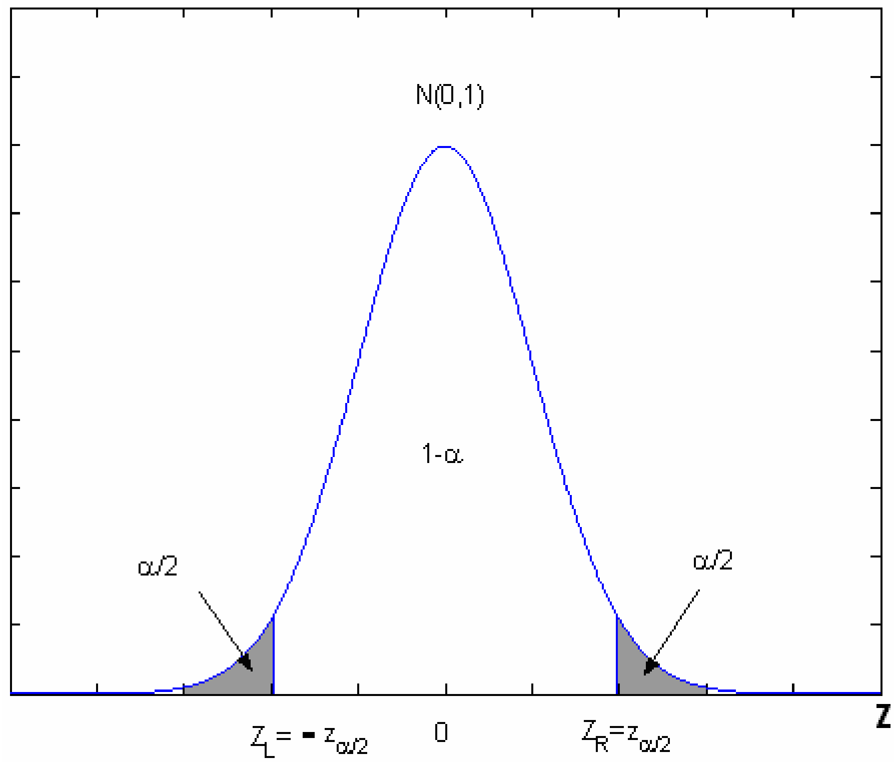
> **Definition.** ***Pivotal Quantity***
> It is a function of the sample and the parameter of interest (μ). Furthermore, we know its distribution entirely.

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$$

Scenario 1 : If σ^2 is known (e.g. $\sigma^2 = 3.7$), then Z is a P.Q. for μ

Scenario 2 : If σ^2 is unknown, then Z is NOT a P.Q. for μ

3. Scenario 1 : Draw the pdf of your PQ.



e.g. Let α = 0.05 in the above figure, we have:

$$P(-Z_{0.025} \le Z \le Z_{0.025}) = 0.95$$

$$P(-Z_{0.025} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le Z_{0.025}) = 0.95$$

$$P(-Z_{0.025} \frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le Z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(-\overline{X} - Z_{0.025} \frac{\sigma}{\sqrt{n}} \le -\mu \le -\overline{X} + Z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(\overline{X} + Z_{0.025} \frac{\sigma}{\sqrt{n}} \ge \mu \ge \overline{X} - Z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

$$P(\overline{X} - Z_{0.025} \frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{0.025} \frac{\sigma}{\sqrt{n}}) = 0.95$$

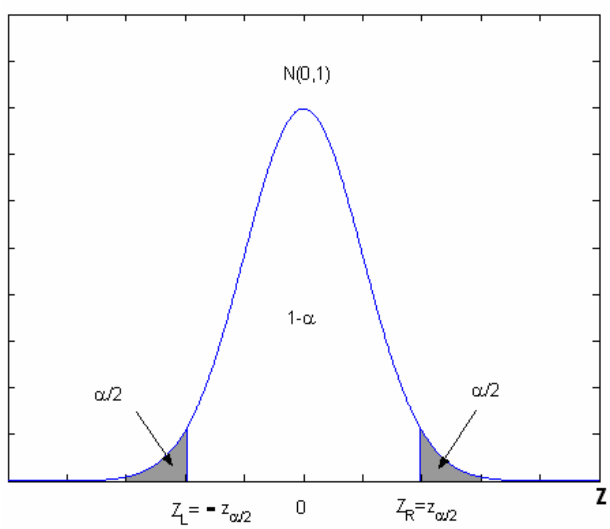
$\therefore$ The 95% confidence interval for μ (when σ^2 is known and the population is normal) is $\left[\overline{X} - Z_{0.025}\frac{\sigma}{\sqrt{n}}, \overline{X} + Z_{0.025}\frac{\sigma}{\sqrt{n}}\right]$

Example. Please calculate the 95% CI for the population mean based on a random sample from a normal population with $\sigma = 3.7,\ \overline{x} = 0.73,\ n = 100$.

Solution. From the Z-table we know that $Z_{0.025} = 1.96$, Therefore the 95% CI for μ is:

$$\left[0.73 - 1.96\frac{3.7}{\sqrt{100}}, 0.73 + 1.96\frac{3.7}{\sqrt{100}}\right]$$

<u>Now we present the derivation of the general formula.</u>



Let α be any small positive value less than 1 (*usually less than 0.5), in the above figure, we have:

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) = 1 - \alpha$$

$$P(-Z_{\alpha/2} \le \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \le Z_{\alpha/2}) = 1 - \alpha$$

$$P(-Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$$P(\overline{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}) = 1 - \alpha$$

$\therefore$ The 100(1-α)% confidence interval for μ (when σ^2 is known and the population is normal) is $\left[\bar{X} - Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right]$

2. (Large Sample) Confidence interval for a population mean (*any population) and a population proportion p

<Thm> Central Limit Theorem

$$\frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} \xrightarrow{n\to\infty} N(0,1)$$

When n is large enough, we have

$$Z = \frac{\bar{X} - E(X)}{\sqrt{Var(\bar{X})}} \dot{\sim} N(0,1)$$

Application #1. Inference on μ when the population distribution is unknown but the sample size is large

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} N(0,1)$$

$$Z = \frac{\bar{X} - \mu}{S/\sqrt{n}} \dot{\sim} N(0,1)$$ (Slutsky's Theorem)

=> $100(1-\alpha)\%$ C.I. for μ $\bar{X} \pm Z_{\alpha/2}\frac{S}{\sqrt{n}}$

Application #2. Inference on one population proportion p when the population is Bernoulli(p)

*** Let $X_i \overset{i.i.d.}{\sim} Bernoulli(p),\ i = 1,\cdots,n$, please find the 100(1-α)% CI for p.

Point estimator : $\hat{p} = \bar{X} = \frac{\sum_{i=1}^{n} X_i}{n}$ (ex. $n = 1000$, $\hat{p} = 0.6$)

Our goal: derive a 100(1-α)% C.I. for p

$$Z = \frac{\bar{X} - E(\bar{X})}{\sqrt{Var(\bar{X})}} \dot{\sim} N(0,1)$$

$$E(\bar{X}) = E(\frac{\sum X_i}{n}) = \frac{1}{n} E(\sum X_i) = \frac{1}{n} \cdot np = p \text{ , } (\because \sum X_i \sim B(n,p))$$

$$Var(\bar{X}) = Var(\frac{\sum X_i}{n}) = \frac{1}{n^2} Var(\sum X_i) = \frac{1}{n^2} np(1-p) = \frac{p(1-p)}{n}$$

$$Z = \frac{\bar{X} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)$$

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \dot{\sim} N(0,1)}$$

$$\boxed{Z = \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \dot{\sim} N(0,1)}$$ (By Slustky's theorem)

$100(1-\alpha)\%$ **(approximate, or large sample) C.I. for p**

$$P(-Z_{\alpha/2} \le Z \le Z_{\alpha/2}) \approx 1-\alpha$$

=> $$P(-Z_{\alpha/2} \le \frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \le Z_{\alpha/2}) \approx 1-\alpha$$

=> $$P(-Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le \hat{p} - p \le Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

=> $$P(-\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \le -p \le -\hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

=> $$P(\hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \ge p \ge \hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}) \approx 1-\alpha$$

=> $P(\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-p)}{n}} \le p \le p + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-p)}{n}}) \approx 1-\alpha$

=> The $100(1-\alpha)\%$ large sample C.I. for p is $[\hat{p} - Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}]$.

CLT => n large usually means $n \ge 30$

special case for the inference on p. n large means

Let $X = \sum_{i=1}^{n} X_i$, large sample means:

$n\hat{p} = X \ge 5$ (X= total # of 'S'), and $n(1-\hat{p}) = n - X \ge 5$ (n-X= total # of 'F')

Example

During one of the "beer wars" in the early 1980's, a taste test between Schlitz and Budweiser was the focus of a TV commercial. 100 people agreed to drink 2 unmarked mugs and indicate which of the two beers they liked better. 54 chose "Bud". Construct and interpret the corresponding 95% confidence interval for p - the proportion of beer drinkers who prefer Bud to Schlitz.

Solution.

<u>Confidence Interval for one population proportion (p) when the sample size is large</u>

Sample size : n $(n = 100)$

Sample proportion : $\hat{p} = \frac{\sum_{i=1}^{n} X_i}{n}$ ($\hat{p} = \frac{54}{100}$)

*** Recall we usually denote $X = \sum_{i=1}^{n} X_i$

"sample is large" means

- For one population mean, $n \ge 30$
- For one population proportion : $X \ge 5$ and $(n - X) \ge 5$

  $(X = 54 \ge 5 \; ; n - X = 46 \ge 5)$

$n = 100,\ X = 54,\ 95\%\ \text{CI for}\ p$

From 95% confidence interval, $1-\alpha = 0.95,\ \alpha = 0.05,\ \frac{\alpha}{2} = 0.025$

$$\hat{p} = \frac{54}{100} = 0.54\ ;\ Z_{0.025} = 1.96$$

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{100}} = 0.049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.049 = 0.096$$

$\therefore$ The 95% confidence interval for p is $[0.444, 0.636]$

If $n = 10000\ ;\ \hat{p} = 0.54$,

$$\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = \sqrt{\frac{(0.54)(0.46)}{10,000}} = 0.0049$$

$$Z_{0.025} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}} = 1.96 \times 0.0049 = 0.0096 \approx 0.01$$

$\therefore$ The 95% confidence interval for p is $[0.53, 0.55]$